1- Qu’est-ce que le Big Data ? : D’où proviennent les données, comment elles sont stockées, dans quels buts sont-elles exploitées ?

L’idée du Big Data (Grande données) est d’enregistrer **le plus de données possible** quelque soit sa forme ou son contenu. L’enregistrement de ses données **augmente d’année en année**, et cela de plus en plus rapidement. Ses données sont enregistré principalement par des **sites internet, les applications mobiles, les conversations téléphonique ou par message**. Mais aussi par les appareils de la vie quotidienne telle que les alarmes incendies, les volets de fenêtres etc…

Ces données vont servir majoritairement au **ciblage publicitaire**, en proposant aux annonceurs de publié des publicités à une catégorie de personne selon les données collectées. Si par exemple une personne est intéressée par le sport, alors des publicités pour des articles de sports vont lui être proposées. Et cela est dû aux recherches qu’il a effectuées auparavant lié à son **centre d’intérêt**.

Ces données ont donc une **finalité économique importante** du fait que les sites récoltes des données pour ensuite les proposés à des annonceurs qui vont pouvoir ciblés des **acheteurs potentiels** qui ont un rapport avec leur activité. Mais aussi ça peut permettre de développer de nouveaux produits en analysant la demande des consommateurs à travers leurs requêtes, tout en dynamisant l’innovation.

On présente souvent le Big Data sous un autre « nom », les **3V**, qui signifient : **V**olume, **V**itesse, **V**ariété. Donc c’est la quantité de données récoltées, stockées puis **exploitées** (analyser), ainsi que la vitesse pour exécuter ses trois étapes, puis la variété des données dû aux types des données qui sont différentes les unes des autres. Il va donc falloir les **structurées et les trier**.

On peut aussi distinguer un autre but de cette **collecte de masse**. Elle va permettre de proposer aux utilisateurs de naviguer sur le web de manière « **gratuite** », comme le dit l’expression « Si c’est gratuit, alors c'est que vous êtes le produit » et donc dans ce cas même ce sont les données qui sont le produit.

2 – Gestion et traitement des données par des algorithmes ainsi que la validation

Le choix des algorithmes et des méthodes, est un **savoir-faire capital** dans le domaine de la data science. L’expertise sur les outils, fait partie intégrante de l’équation. De même, penser le modèle avec l’**industrialisation** en tête est un point clé de succès.

<u>Le choix des algorithmes est donc fait en fonction d’un grand nombre de paramètres :</u>

• La **qualité et la disponibilité** des données en entrée qui sont en amont traitées et collectées

• Les **contraintes d’industrialisation** (comme par exemple le temps de calcul)

• La **vitesse d’exécution** de l’algorithme soit précis et robuste

• Les **outils** qui doivent être utilisés pour que l’algorithme soit performant et fiable

• Le **type d’infrastructure** à mettre à disposition (serveur de fichiers par exemple)

Avant toute mise en production, il est également important que les modèles établis soient testés avec le plus grand soin.

Cette évaluation se fait (pour rappel) principalement sur deux critères qui sont :

• La **précision** : pour que l’algorithme soit valide, il faut qu’il soit **précis dans ses recherches** et donc dans ses résultats, pour offrir à l’utilisateur des **résultats fiable**.

• La **robustesse** : l’algorithme doit être robuste, c’est-à-dire qu’il doit permettre d’**exécuter toutes les requêtes effectuées** par l’utilisateur, on va donc parler d’algorithme efficace.

La précision dépend principalement de la **taille de l’échantillon** et de la **complexité** du modèle. Plus le modèle est complexe, plus il est précis mais plus on risque un **sur-apprentissage**, particulièrement si la taille de l’échantillon est petite. C’est pourquoi il est indispensable de tester la robustesse qui permet de garantir que le modèle, une fois mis en production sur des données réelles, ne perdra pas trop en précision.

3 – Ethique de cette pratique

Il y a l’émergence évidente de **problèmes éthiques**, comme par exemple l’**exploitation des données médicales**. Nous pouvons prendre l’exemple de l’hôpital de Doullens où l’utilisation d’un algorithme pose problème sur le fait que le stockage des **données sensibles** ainsi que son utilisation. Il faut donc dans ce cas faire en sorte de **sécurisé les données stockées**. De même, pour que l’algorithme soit valide, donc exploitable, cette collecte doit être éthique, car les données médicales sont personnelles et relèves du **secret médical**, donc une responsabilité **juridique** et **éthique**.

Sources :

L'essentiel du Big Data : http://www.cea.fr/comprendre/Pages/nouvelles-technologies/l-essentiel-sur-le-big-data.aspx

Comprendre ce que permet le Big Data : https://www.culture-informatique.net/cest-quoi-le-big-data/

Introduction d'algorithme dans un hôpital : https://www.liberation.fr/france/2018/04/29/a-l-hopital-de-doullens-big-debat-surles-big-data_1646748